Mathematical Programming manuscript No.
(will be inserted by the editor)

Elizabeth Eskow · Brett Bader · Richard Byrd · Silvia Crivelli · Teresa Head-Gordon · Vincent Lamberti · Robert Schnabel

An optimization approach to the problem of protein structure prediction

Received: date / Revised version: date

Abstract. We describe a large-scale, stochastic-perturbation global optimization algorithm used for determining the structure of proteins. The method incorporates secondary structure predictions (which describe the more basic elements of the protein structure) into the starting structures, and thereafter minimizes using a purely physics-based energy model. Results show this method to be particularly successful on protein targets where structural information from similar proteins is unavailable, i.e., the most difficult targets for most protein structure prediction methods. Our best result to date is on a protein target containing over 4000 atoms and $\sim$12,000 cartesian coordinates.

1. Introduction

The protein structure prediction problem is one of the fundamental challenges of modern science. It is to predict the three-dimensional shape, or native state of a protein, given its sequence of amino acids. Optimization is one of the promising approaches to solving this problem because it is believed that, in most cases, the native state corresponds to the minimum free energy of the protein. However, the energy landscape of a realistic-sized protein has thousands of parameters and an enormous number of local minimizers. This means that an efficient global optimization approach for very large scale problems is required to solve the problem.

The global optimization method discussed in this paper is able to make progress on large-scale problems, such as proteins with over 200 amino acids and thousands of cartesian coordinates, by performing global optimization on small-scale subproblems and local optimization on the full dimensional parameter space. Two features related to the specific class of problems also are integral to its success. First, using secondary structure predictions to configure the more basic portions of the protein is an important feature of the method. This limits the number of candidate configurations for the protein target, which is known to be exponential in the number of amino acids. Second, an optimization approach will only succeed if the energy model is sufficiently accurate, and this model must take into account the solvation environment of proteins. Our approach uses a

B. Bader, R. Byrd, E. Eskow, V. Lamberti, R. Schnabel: Department of Computer Science, University of Colorado at Boulder, Boulder, Colorado 80309-0430

S. Crivelli, T. Head-Gordon: Physical Sciences, Life Sciences and NERSC Divisions, Lawrence Berkeley National Laboratory, Berkeley, California 94720

unique model of solvation in conjunction with one of the well-known empirical energy potentials.

In Section 2, we describe the types of methods that are generally used for protein structure prediction. The energy model, including our new solvation term, is described in Section 3. A detailed description of our two-phased global optimization approach is given in section 4, followed by results and discussion in section 5. These include promising results from a recent, blind comparison of protein structure prediction methods. Section 6 contains a summary and conclusions.

2. Background information and approaches to protein structure prediction

2.1. Background Chemistry Information

To understand the description of this global optimization method compared to other protein structure prediction methods, a little background about proteins is necessary. The basic building blocks of proteins are amino acids. The amino acids are bonded together by peptide bonds to form a polypeptide chain, and the ordered sequence of amino acids in the chain is referred to as the primary structure of the protein. For a given protein target, this sequence is known and is the starting point for the protein structure prediction problem. The secondary structure of a protein refers to regular, local structure in portions of the polypeptide chain, such as α-helices, β-sheets and the turns that link them. Tertiary structure describes the overall shape adopted by the polypeptide chain. The goal of protein structure prediction is to find the tertiary structure of a target protein given its primary structure of amino acids.

2.2. Non-optimization Approaches

There are two major types of non-optimization methods for protein structure prediction, both of which involve comparisons to known protein structures from a protein database. Comparative modeling methods [3,44,14,4,15] make these comparisons based on finding similarities in the primary sequence of amino acids, whereas fold recognition methods [21,13,22,23] try to apply known standard folds to the target protein. The most successful methods at the recent CASP4 conference (Fourth Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction, December, 2000) use information from the sequence and structure of known proteins to form templates for predicting tertiary structure of unknown targets. For targets where this information is unavailable, these methods may be somewhat less successful and optimization approaches may be particularly important.

2.3. Optimization Approaches

Optimization approaches strive to find the protein conformation that minimizes the energy of a rugged energy landscape, a difficult global optimization problem with expontially many local minima and a huge parameter space. This approach should be capable of finding protein structures that are very different from known structures. Because of the difficulty of the optimization problem many of the methods of this time make some use of protein database information. Some structure prediction methods use optimization techniques only for those portions of protein targets whose sequence or structure do not exist in current databases because they are "new folds" which cannot be recognized with information from the set of currently known proteins. At the opposite end from the non-optimization approaches described above, the method of Harold Scheraga's group ([28, 27]) uses the information of known protein structures only in determining the weights of various terms in their simplified potential energy function. After preliminary global optimization on this simplified function a Monte Carlo-based optimization is performed with a physics-based energy for the final prediction. The approach to protein structure prediction described in this paper uses information from known proteins only to make secondary structure predictions, but not in the tertiary structure predictions or in generating the terms of the physics-based energy function used.

3. The Energy Function

This work and almost all other work in optimization approaches to molecular structure prediction is based on the use of empirical energy potentials, which are fairly simple mathematical formulas that are reasonable approximations to the potential energy for particular classes of molecules. Commonly used empirical potentials include CHARMM [5], AMBER [9] and ECEPP [31]. The AMBER molecular mechanics energy function, E_{AMBER}, like many other functions represents the cartesian coordinates of the n atoms of the protein as a vector of length $3n$. The positions of atoms may also be described by parameters r_i, the distance between the i-th atom and a designated neighboring atom, θ_ℓ the bond angle formed by a sequence of three bonded atoms, and ϕ_k a dihedral angle formed by a sequence of four bonded atoms. These parameters are used in part of E_{AMBER}, which has the form

$$E_{\mathrm{AMBER}} = \sum_{\text{bonds}} K_{r_i}(r_i - r_{i,eq})^2 + \sum_{\text{angles}} K_{\theta_\ell}(\theta_\ell - \theta_{\ell,eq})^2 + \sum_{\text{dihedrals}} \frac{V_k}{2}[1 + \cos(n_k\phi_k - \gamma_k)] + \sum_{i<j}\left(\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + C\frac{q_iq_j}{r_{ij}}\right) \tag{1}$$

The first 3 terms represent the bonded interactions, and the final terms represent the non-bonded interactions, which are comprised of Lennard-Jones and electrostatic terms, respectively. These non-bonded interactions occur between every pair (i, j) of atoms and depend on the distance r_{ij} between the pair of atoms, and on their charges $q_i q_j$. The non-bonded interactions are targeted in our "smoothing" approach described in (Sect. 5.4.1) because their extreme nonlinearity influences the difficulty of the minimization problem.

A crucial component of our energy model is the development of additional terms to the energy function that more accurately model solvation effects, i.e. the behavior of the protein in an aqueous environment. It has been shown clearly by various research groups that earlier potentials do not adequately model solvation effects and that this shortcoming significantly hinders the ability to utilize them to determine structures of proteins by optimization-based approaches [26,32]. In particular, new terms are needed that model hydrophobic (adverse to interaction with water) and hydrophilic behavior.

To address this issue, an empirical solvation free energy term $E_{\mathrm{SOLVATION}}$ has been added to the energy function used by the optimization. $E_{\mathrm{SOLVATION}}$ models the hydrophobic effects as a two-body interaction between all aliphatic carbon (carbon atoms not bonded to electronegative atoms, such as oxygen) centers. This description is motivated by recent experimental, theoretical and simulation work to determine the role of hydration forces in the structure determination of model protein systems [33,18,40,19,39]. This work and that of others on the free energy of association of two small hydrophobic groups (of methane or leucine molecules) in water show this interaction potential has two minima separated by a barrier: one for the molecules in contact and one for the molecules separated by a distance of one water molecule. Our new solvation term embodies these characteristics. Using a sum of Gaussians

$$E_{\mathrm{SOLVATION}} = \sum_{i,j \leq N_c} \sum_{k \leq M} h_k \exp\left(- \left[\frac{(r_{ij} - c_k)}{w_k} \right]^2 \right), \tag{2}$$

where the sum over i and j is over the aliphatic carbon centers, and each of the M Gaussians is parameterized by position (c_k), depth (h_k), and width (w_k) so as to describe the minima and barrier. The benefits of this form are that (1) it introduces a stabilizing force for forming hydrophobic cores, (2) it is a well-defined model of the hydrophobic effect for hydrophobic groups in water, and (3) it can be described as a continuous potential that is much more computationally tractable than an alternate approach, solvent accessible surface area models.

We tested the effect of the solvation energy function in the form (2) on conformations of a 70 amino acid protein uteroglobin (2utg_A) and a 71 amino acid DNA binding protein (1pou). Using this form of potential, we found good agreement with experiment in that the potential energy of the experimentally determined structure was lower than the potential energy for any of the structures found by the global optimization algorithm. In order to get such good agreement we used parameter values in (2) that exaggerate the stability of the contact and solvent-separated minima in comparison to the parameter values based on the experimental methane work mentioned above.

4. The global optimization approach to protein structure prediction

The stochastic-perturbation global optimization method for protein structure prediction (henceforth referred to as SPGOPS) consists of two phases. The first phase generates a set of good initial structures using information specific to the domain of protein structure prediction. This information is the prediction of secondary structure, which is readily available through servers (e.g. [20,12]), with accuracies averaging around 75%. Given the widespread use and high accuracy of secondary structure prediction methods, it would be impractical not to attempt to utilize them in this problem domain and only by using them can we realistically tackle reasonable-sized problems. Secondary structure prediction is incorporated into the first phase with the goal that the second phase begins with structures that have much of the predicted secondary structure. This is accomplished with the use of "biasing" functions, which are penalty terms that bias the structures towards predicted secondary structure [16], and are described below.

The second phase improves upon the initial structures by a combination of breadth (work on a variety of structures) and depth (improve the current lowest energy structures). It does this by a combination of steps that change the structure in a restricted manner – this is accomplished by small scale global optimization using a stochastic method based on [36] applied to a selected small subset of the parameters – and local minimizations applied to the full structure. A key to this step is selecting a small subset of parameters that, by changing their value, can lead to a substantially improved structure for the entire problem. Heuristics for doing this are discussed in section 4.2. A list is kept of all minima obtained from the local minimizations, and after some (fixed) number of iterations, this list is clustered via pairwise root mean squared deviation (RMSD) evaluations, and ordered by energy value within each cluster. The RMSD of 2 configurations measures the closeness of their structures, i.e. a few Å difference means the structures are very similar, and beyond 10Å usually means they are not. After clustering, if the stopping criteria described below have not been met, additional iterations of the second phase are applied, starting from a new list containing the lowest energy minimizer from each cluster.

A very general framework that shows how the parts of the global optimization approach fit together is given in Algorithm 4.1.

4.1. Phase 1: Generation of Starting Configurations

Phase 1 builds initial configurations that contain predicted secondary structure. To accomplish this the "antlion" method [17,16] is used to apply server-predicted secondary structure information in energy minimizations. The server predicts whether each amino-acid of the target protein is α, β or coil (i.e. not α or β), and gives a strength for each prediction based on tendencies from other known proteins. The "antlion" strategy uses biasing or penalty functions designed to enforce the information from the secondary structure prediction server in step(1b)

Algorithm 4.1 – Basic Framework of SPGOPS

1. **Phase 1: Coarse Identification of Configurations :**
   Generate initial configurations containing domain specific tendency information.
   (a) **Sampling in Full Domain:**
   Generate the parameters of some sample configurations.
   (b) **"Biased" Full-Dimensional Local Minimizations :**
   Use server obtained information to create biasing terms for predicted secondary structure. Perform a local minimization from a subset of the sample points, using "biasing" penalty functions to superimpose predicted secondary structure on the standard energy surface.
   (c) **"Unbiased" Full-Dimensional Local Minimizations :**
   Perform a local minimization from each "biased minimizer" using the "unbiased" energy function. Save these local minimizers for Step 2a.
2. **Phase 2: Improvement of Local Minimizers:**
   For some number of iterations:
   (a) **Select a Configuration and Small Subset of Parameters to Improve:**
   From the list of full-dimensional local minimizers, select a local minimizer to improve. Then select a subset of the parameters of this configuration to optimize in step 2b.
   (b) **Small Sub-problem Global Optimization:**
   Apply a global optimization algorithm to the energy of the selected configuration with only the selected parameters as variables.
   (c) **Full-Dimensional Local Minimization :**
   Apply a local minimization procedure, with all the parameters as variables, to the lowest energy configurations that resulted from step 2b, and merge the new local minimizers into the list of local minimizers.
   (d) **Cluster Local Mimima and Test for Convergence**
   Cluster the list of local minimizers via pairwise RMSD and if the stopping criteria has not been met, repeat all steps of Phase 2 from a new list of local minimizers containing the lowest energy minimizer from each cluster.

of Algorithm 4.1. The biasing functions used are described in the two subsections below.

To begin the entire process, a buildup procedure is used. This procedure samples on the set of dihedral angles for each amino acid which has been predicted to be "coil" some fixed number of times, and selects the angle values that produce the best partial energy for the part of the chain built so far, before proceeding to the next amino acid. For amino acids which are predicted to be α or β, the dihedral angels are set to the ideal values for those types structures. A subset of the best structures generated by this buildup procedure is then selected as starting points for full dimensional local minimizations using the antlion strategy, and some number of the best minimizers generated are passed to the second phase of the algorithm.

4.1.1. Biasing functions for α-helical Proteins α-helices are rodlike contiguous segments of tightly coiled polypeptide chain. The α-helix is stablized by hydrogen bonds connecting the carbonyl oxygen of residue i to the amide hydrogen of residue $i + 4$. Two functions have have proved very effective in encouraging formation of α-helices. The first function,

$$E_{\phi\psi} = \sum_{\text{dihedrals}} k_\phi[1 - \cos(\phi - \phi_0)] + k_\psi[1 - \cos(\psi - \psi_0)] \qquad (3)$$

biases the backbone torsional angles of the amino acids of a residue that is predicted to be an α-helix to be close to ideal values for an α-helix. Here ϕ_0 and ψ_0 are the dihedral angles of a perfect α-helix, and k_ϕ and k_ψ are force constants related to the strength of the prediction from the prediction server. The second function,

$$E_{HB_\alpha} = -w_{i,i+4}/r_{i,i+4}. \tag{4}$$

encourages α-helical hydrogen bonds to form between the oxygen of residue i and the hydrogen of residue $i+4$, if residues i and $i+4$ are predicted to be helical. In this function, $w_{i,i+4}$ is the weight of the prediction confidence output by the server, and provides a strong incentive for an intramolecular hydrogen bond to form when residues i and $i+4$ are strongly predicted to be helical. Together these functions "bias" the protein towards forming α-helical shape in regions predicted to be helix.

4.1.2. Biasing functions for Proteins containing β-sheets A β-sheet is comprised of 2 or more non contiguous strands of contiguous amino acids with hydrogen bonds connecting the carbonyl oxygen of one strand to the amide hydrogen of another. The two strands that are connected in a β-sheet may run in the same direction along the chain (parallel) or in the opposite direction (antiparallel), and they may be any distance apart on the amino acid chain. Forming β-sheets in proteins is an intrinsically hard problem. Since α-helices contain hydrogen bonds along the backbone from neighboring amino acids, these interactions are relatively short and local, spanning only four residues. On the other hand, β-sheets usually have nonlocal hydrogen bonds, where the hydrogen bonds span many more residues. This nonlocal nature of β-sheets requires a modified approach to form them in a protein structure prediction algorithm. Secondary structure predictions provide information regarding which residues or segments of the amino-acid chain are predicted to be β-strands, but not how these strands align to form β-sheets.

When using secondary structure predictions of β-strands, the challenge is to correctly pair and align the β-strands into the correct β-sheet(s) in the tertiary structure. One of the difficulties that must be overcome is the combinatorial explosion of possible β-strand matches. With only two strands identified, the choices are limited to parallel versus antiparallel orientations and hydrogen bonds on the even or odd residues. Two strands of equal length produce four possible matches. Often, the predictions do not yield strands of equal length, which further multiplies the combinations by the offset length plus one. The presence of additional strands complicates matters further by introducing even more possible matchings, and the problem grows exponentially harder with each additional β-strand. Thus, a biasing scheme that tests each possibility in turn would require many time-consuming runs.

We limit this combinatorial complexity by removing some of the potential matchings via a preprocessing step. The technique used by SPGOPS is to evaluate each possible pair of strands using a scoring function based on occurrences of β-sheets in a protein database ([47]). The scoring function returns a value representing the bonding probability of a pair of residues for forming β-sheet type hydrogen bonds. The scores of each possible pair of β-strands, with varying alignments, are summed, and the best-scoring, physically feasible set of strand pairs is selected for use in the β biasing function given below. If more than one set of good-scoring, feasible strand pairs is identified, each is used in a separate instantiation of Phase 1. Results from the various runs of Phase 1, each using a different set of strand pairings in the biasing function, are compared and the best structures are selected by energy values and number of contacts (hydrogen bonds between β-strands) formed.

Since the set of hydrogen-oxygen pairs along the two strands includes both the H-bonded and non-H-bonded pairs, a biasing function that handles both types concurrently and without any explicit identification is necessary. In our work, this is accomplished by the following piecewise continuous biasing function, which couples a linear function with a sigmoid function,

$$E_{HB_\beta} = \begin{cases} \frac{\epsilon\,\tau}{4}\left(\frac{r_{ij}-\sigma}{\omega}+2\right) & \text{if } r_{ij} > \sigma \\ \frac{\epsilon\,\tau}{1+\exp(\frac{\sigma-r_{ij}}{\omega})} & \text{if } r_{ij} \leq \sigma \end{cases} \tag{5}$$

where ϵ is the dielectric constant, r_{ij} is the Euclidean distance between atom i and atom j, τ is a scale factor for appropriate balance with other forces in the model, σ is the sigmoid offset from the origin, and ω scales the sigmoid width. The linear term ($r_{ij} > \sigma$) attracts atom pairs from afar to be at least the distance of a typical non-H-bonded pair. Then the attractive force within the sigmoid term ($r_{ij} \leq \sigma$) decays as the two atoms move nearer to each other. There is still enough attraction, however, for the biasing function in conjunction with the other terms in the energy function to encourage a hydrogen bond to form between the most likely H-bonded pair, yet not too much force to disrupt a non-H-bonded pair. The three parameters in the biasing function (τ, σ, ω) affect the formation of hydrogen bonds in β-sheets. Our current slate of experimental parameters ($\tau = 2.09$, $\sigma = 16$, $\omega = 4$) appears to work well, but we still believe there is room to improve this biasing function by tuning these parameters. In addition to the biasing term (5) we also a use a biasing term analogous to (3) biasing the torsion angles along the predicted β-strand towards ideal values for a $beta$-sheet.

4.2. Phase 2: Improvement of local minimizers

The second phase accounts for most of the computational effort of the method. The basic idea of this phase is to select a promising configuration from the list of local minimizers, and then select a small subset of its variables for improvement followed by full variable local minimization from the best resultant configurations.

To describe the strategy for selecting configurations, we can use the model of tree searching, where a tree consists of an initial minimizer and all the minimizers generated from it so far by applying a global optimization on a small subspace of torsion angles followed by a local minimization over the entire space. In the initial iterations (typically ~10) of Phase 2, the tree with the least amount of work performed on its members so far is selected, and the lowest energy configuration in this tree that has not already been used is chosen as the configuration to be improved. After the fixed number of iterations of this "balancing" stage, the remaining iterations (also typically ~10) of Phase 2 correspond to the "non-balancing" stage. In this stage, the lowest energy configuration is selected, regardless of which tree it comes from. We have found that the combination of breadth and depth in the search of the configuration space contributes to the success of this method.

Once a configuration is chosen, a small subset of its variables is chosen for modification by global optimization. The subset of variables consists of a small number of dihedral or torsional angles of the protein picked randomly from the set of amino acids predicted to be "coil" by the secondary structure prediction. Once the subset has been determined, a stochastic global optimization procedure similar to the one in [36] is executed to find the best new positions for the chosen dihedral angles, while holding the remaining dihedral angles fixed. This stochastic global optimization approach provides a theoretical guarantee of finding the global optimum, and tests have shown it to be efficient compared to other global optimization approaches for problems with small numbers of parameters [37]. The global optimization method samples over the parameter space, and it selects a sample point to be a start point for a local minimization only if that sample point has the lowest energy of all neighboring points within a certain distance metric. If a sample point lies within the distance metric of another point that is lower in energy, it is assumed to lie within an existing basin of attraction and is rejected from further computational consideration. Because the probabilistic theoretical guarantee is easier to satisfy computationally for small dimensional problems, we select a subset size of ~6-10 variables (from the space of torsion angles of the protein). This global optimization algorithm is general in the sense that a parameter space of arbitrary dimension can be explored, however, in practice, the amount of work required to reach the theoretical guarantee is prohibitive for large subspaces.

The small-scale global optimization allows the method to explore configurations with significantly different shapes than the ones that it starts with. It produces a number of local minimizers in the small subspace of chosen parameters (torsion angles). A number of those conformations with the lowest energy values are selected for local minimizations (or "polished") in the full variable space. These full-scale local minimizations are less likely to produce major structural changes but can cause important, more local refinements throughout the protein. The new full-dimensional local minimizers are then merged with those found previously, and the entire phase is repeated a fixed number of times.

Periodically during Phase 2 the minima are ordered and clustered to decide whether to terminate Phase 2 or continue. The clusters are defined so that members of each cluster are within ~5 Å RMSD of the lowest energy configuration in that cluster. The number of clusters indicates the number of diverse structures that exist at this stage. If the energy of the global minimum is no longer decreasing, then the algorithm is considered to have converged. Conversely, if the energy of the global minimizer is continuing to decrease, then more iterations of Phase 2 are performed. In that case, the list of minimizers is reduced to the set consisting of the lowest energy minima from each cluster.

5. Results and Discussion

The results of the SPGOPS approach appear to be at the leading edge of protein structure prediction via optimization using a physics-based energy function, along with a small number of other groups such as [28]. In this section, we present results of SPGOPS on 2 helical proteins, and some results from the recent CASP4 experiment in which 163 research groups worldwide participated in structure prediction of 43 target proteins whose structures were unknown for the duration of the experiment and only recently released to CASP4 participants. We highlight some of our CASP4 results to demonstrate the progress of the optimization approach in predicting the structures of more difficult protein targets. All of our CASP4 results will be presented and analyzed in [10].

5.1. Prior Results on Helical Proteins

Recent research with our approach involved the prediction of the structure of helical proteins with about 70 amino acids [11]. Starting with predictions of secondary structure from the neural network package of [45] that are incorporated in Phase 1 of SPGOPS (described in Sect. 4.1), this method has determined the structure of two helical proteins with 70 and 71 amino acids, the A-chain of uteroglobin, 2utg_A, and a four-helix bundle DNA binding protein, 1pou. Resulting predictions for the entire structures are within 7.3 Å RMSD of the crystal structure for uteroglobin and 6.3 Å RMSD of the NMR structure for 1pou. Substituting correct (known) secondary structure for the secondary structure predictions within the global optimization for 1pou resulted in structures with RMSD's of under 5 Å. According to a recent article by [35], the probability of obtaining a 6 Å RMSD by chance is so remote that a prediction algorithm obtaining such structures should be considered quite successful. These results show good ability of our method to predict the structure of moderate-sized α-helical proteins, and also demonstrate that as secondary structure prediction improves, the tertiary structures produced by our method will become more accurate as well.

5.2. Recent Results from CASP4

The CASP4 experiment ran from May 11 through Sept 15, 2000. During this time the amino acid sequences for 43 target proteins, whose structures were just in the process of being determined experimentally, were released to participants for "blind" prediction. The targets were released and predictions due on various

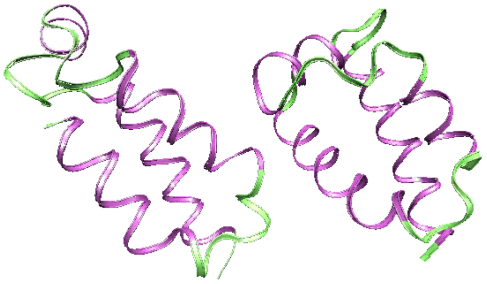

Fig. 1. Results for SPGOPS on 1pou with RMSD of 6.3 Å. The NMR structure is on the left, and the SPGOPS result on the right.

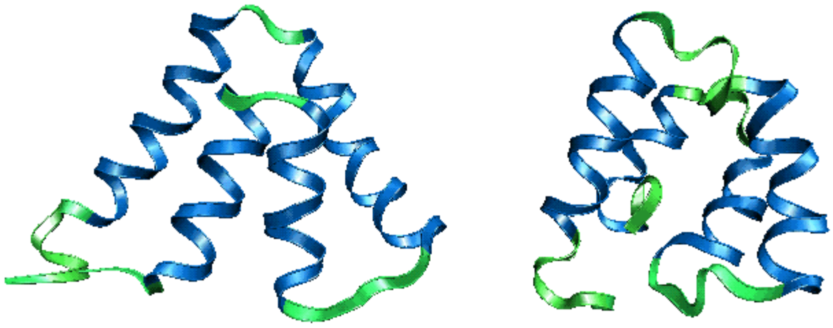

Fig. 2. Results for SPGOPS on uteroglobin with RMSD of 7.3 Å. The crystal structure is on the left, and the SPGOPS result is on the right.

dates during that time period, and the amount of time allocated for predicting a specific target also was variable. Up to five models of structure prediction for each target were accepted, however, the submission specified as “model 1” was predominantly used for evaluation of structure predictions. We submitted models for 8 different target proteins, sometimes submitting more than 1 model per target, but will only discuss our “model 1” results here. We always chose as “model 1” the prediction with lowest potential energy. First we present results on secondary structure prediction accuracy for all 8 targets, showing the overall effectiveness of Phase 1 of SPGOPS. Then we discuss results for the 8 targets compared to the other CASP4 submissions, relative to the difficulty of predicting the targets based on the criteria used in CASP4. Finally, we present performance results for two of our target submissions: a 242 amino acid α-helical protein and a 56 amino acid β protein.

5.2.1. Secondary Structure prediction accuracy Secondary structure prediction methods in general are much more advanced than methods predicting overall tertiary structure, and we take advantage of them in the SPGOPS method. While forming predicted α-helical segments using the biasing functions described in section 4.1 is not very difficult due to the local nature of the hydrogen bonds in α-helices, this task is much more difficult in the case of β-sheet formation. β-strands may be located at distant parts of the protein, and there may be only 1 or 2 hydrogen bonds formed to hold β-strands together in a β-sheet. We had not used our β biasing techniques in their current form prior to CASP4. Since we attempted to predict four CASP4 targets with β-sheets that we were able to form to some degree, CASP4 proved to be a good test for our new biasing techniques.

Figure 3 shows 3 lines of secondary structure for each target consisting of (1) the secondary structure predictions we used (from [20], [12] or some combination of prediction servers), (2) the secondary structure in our submitted model 1, and (3) the secondary structure of the target protein. The darker lines in the figure represent helical segments and the light segments are β strands. For targets 91, 99, 110 and 124, some parts of the secondary structure of our models are closer to the target's actual secondary structure than what was predicted. This shows that our method not only can incorporate predicted secondary structure (in Phase 1) but sometimes can improve upon it significantly (in Phase 2). On the other hand, for targets 97, 105 and 125, the predicted secondary structure was closer to the target's actual secondary structure than the secondary structure obtained in our models although this deterioration is not nearly as significant as the improvements in the previous set of cases. Target 98 had the predicted secondary structure implemented accurately, but the secondary structure prediction was a bit different from the target secondary structure.

The overall secondary structure accuracy for the model we submitted for each target is evaluated by two numbers which are given to the left of the figure. The "Q3" percentage is measured by the percentage of helical, beta and coil residues predicted correctly over the number of all residues of the protein. The segment overlap measure (SOV) is a more structurally meaningful measure of secondary structure prediction accuracy that also ranges from 0 to 100, and is defined in [46]. For target 105, the SPGOPS method did not form the predicted β-sheet, and although the individual β-strands in the model are not very far off from being within hydrogen bond distance, the Q3 and SOV scores are both very poor for that target. Except for Target 105, the Q3 measures of our models range from 73.1 to 93.0, and the SOV measures range from 68.9 to 92.6, showing that the incorporation of predicted secondary structure into the submitted models, via the initial configurations generated by Phase 1 of SPGOPS is reasonably successful. It is also interesting to note that in some cases where the predicted secondary structure was in error, our algorithm was able to correct this error to some extent in making its prediction.

5.2.2. Tertiary structure prediction performance The organizers of CASP4 have ranked all of the targets with respect to sequence homology and fold recognition, such that the "easiest" targets have a high percentage of their sequence similar to known proteins, "harder" targets have some but less structural sim-

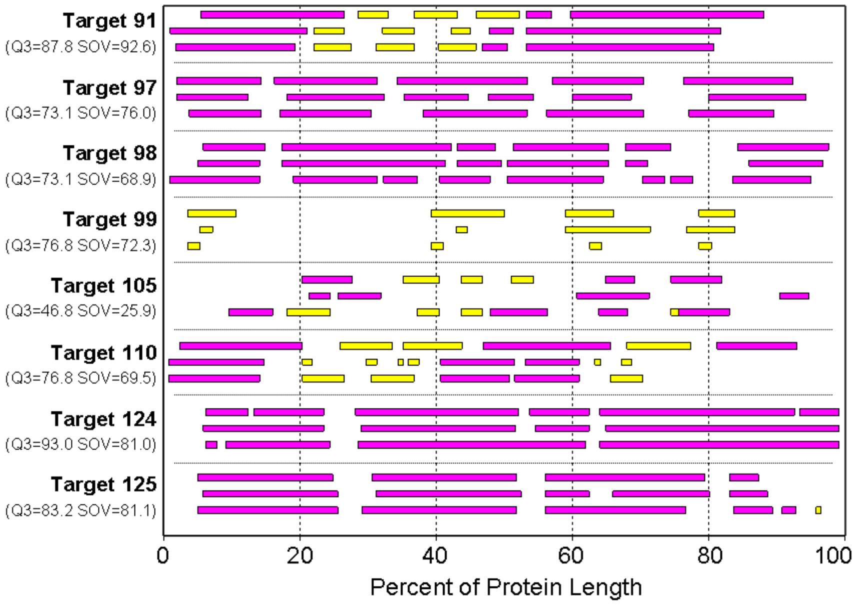

Fig. 3. The 3 lines for each target represent (1) the secondary structure predictions used to generate initial configurations in Phase 1 of SPGOPS, (2) the secondary structure of our model 1 submission, and (3) the secondary structure of the experimental structure. The dark lines denote α-helical and the lighter lines are β-strands. To the left, "Q3" and "SOV" are measures of the percentage of secondary structure accuracy of our model 1 submissions.

ilarity to known proteins, and the hardest targets have very little similarity to known proteins and are called "new folds". The CASP4 organizers classified all 43 targets into 8 difficulty bins. In figure 4, we plot the comparative difficulty of each protein we attempted in CASP4 versus the comparative accuracy of our prediction to that of other groups. The difficulty measure for each target is its (CASP4 bin number)/8 *100. The comparative accuracy measure is based on the overall accuracy measure used in CASP, GDT_{TS}. GDT is the global distance test which measures the largest set of contiguous residues whose RMSD from the target is under a certain distance cutoff. The measure GDT_{TS} is the GDT total score, measured as

$$GDT_{TS} = (GDT_{P1} + GDT_{P2} + GDT_{P4} + GDT_{P8})/4.0, \tag{6}$$

where each GDT_{Pn} is an estimate of the percent of residues under distance cutoff $\leq n.0$ Å. The comparative accuracy ranking is the percent of groups with poorer GDT_{TS} scores on that protein.

The two plots in Figure 4 show that as the difficulty of the targets increases, the GDT_{TS} percentile of our models ranked against all other model 1 submissions generally increases as well. In other words, the SPGOPS method is relatively more effective on targets where less information from known proteins is available. This is true because most methods used for CASP4 predictions relied heavily on knowledge from known proteins, whereas SPGOPS uses that knowledge only in forming secondary structure, but not in the prediction of overall tertiary structure. This shows that our approach has potential value where other approaches have limitations.

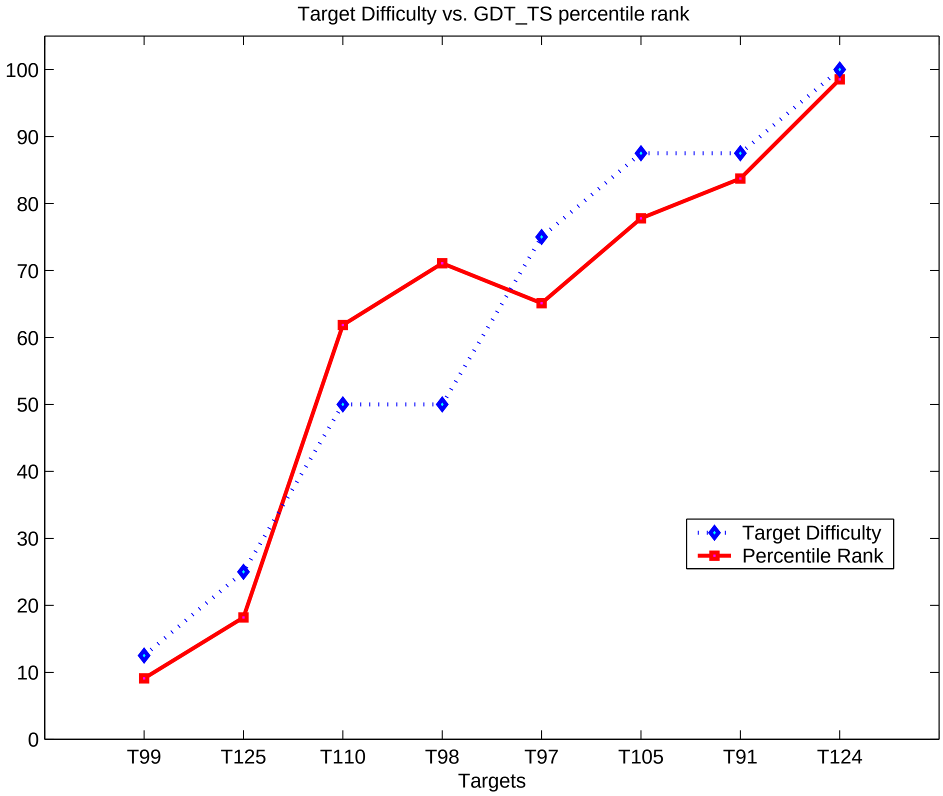

Fig. 4. Difficulty of CASP4 targets as rated by CASP4 organizers (— — ——) versus the percentile ranking of our model 1 submissions (——) using SPGOPS. The percentile ranking of our models generally increases with target difficulty.

Two of the targets we submitted for CASP4 are interesting for very different reasons. Our submission for Target 99 (a synthetic construct) was relatively poor in comparison to other predictions because this target had high sequence homology, meaning that it closely matched known proteins. However, this is a beta sheet protein, and while we didn't get the overall topology right, without using any knowledge of sequence homology we predicted a good portion (30/54 amino acid fragment shown in Figure 6) with an RMSD of under 5 Å. The overall RMSD for our target 99 model 1 submission is 7.79 Å (see Figure 5).

Target 124 (Phospholipase C beta C-terminus, turkey) was considered to be a hard target, or new fold, by the CASP4 organizers. It was also a difficult target from an optimization point of view, with 242 amino acids, 4102 atoms, and over 12,000 cartesian coordinates. Our model 1 submission (Figure 7) was among the best predictions for this target, with the best GDT_{TS} score, and an overall RMSD of 8.46 Å. After the deadline for submission to CASP4, an additional run of Phase 2 was performed on our CASP4 result for Target 124. This new run lowered the energy of the global minimizer from -4528 to -5095, resulting in a new RMSD of 7.7 Å. The results for this target again show the value of a physics-based optimization method that does not rely on known protein structures for predicting proteins with new folds.

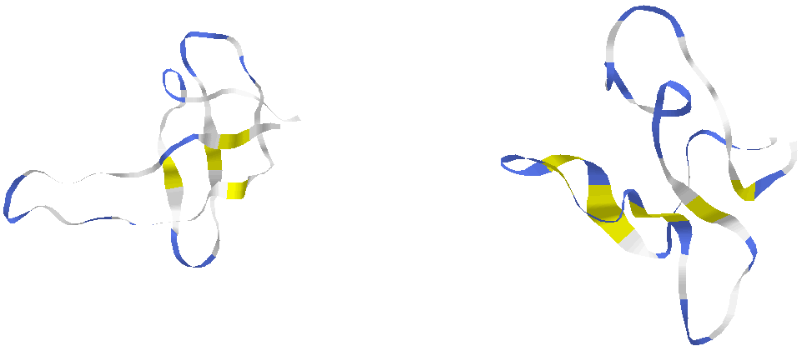

Fig. 5. Results for SPGOPS on Target 99 (containing 56 amino acids) with RMSD of 7.79 Å, the right structure is our lowest energy submission, the left is the NMR structure.

5.3. Computational costs

In order to give some indication of the cost of using SPGOPS, we give an estimate of the run times and amounts of computation for each of the two CASP4 predicted targets (99 and 124) highlighted in the previous subsection. For target 99, Phase 1 (step (1a)) of Algorithm 4.1 generated 10 starting configurations. A local minimization of ~12,000 steps using beta-biasing (step 1b) was applied to each. After Phase 1, the 10 initial minima were clustered, resulting in 3 diverse clusters of which the lowest energy minima from each was passed on to Phase 2. A total of 21 iterations of phase 2 were computed, 9 of which used the balancing paradigm for choosing configurations to minimize and 12 used the nonbalancing paradigm. The structure we submitted to CASP4 had the lowest overall energy, and the total computation time was roughly 36 hours elapsed

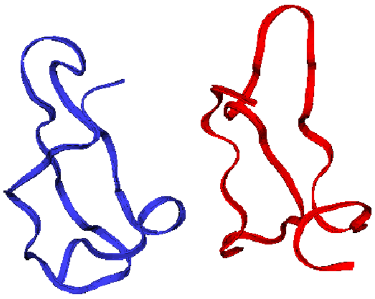

Fig. 6. Results for SPGOPS on Target 99 for a 30 amino acid fragment with RMSD of 4.95 Å, the right structure is our lowest energy submission, the left is the NMR structure.

time using 10 Avalanche (300 mhz) workstations at the University of Colorado. For target 124, Phase 1 used only 1 extended conformer as the starting configuration, and local minimization with alpha-biasing was done in portions over different segments because the configuration was too long to bias all the helix at once. It took $\sim$25,000 iterations for all of the α-helices to form. Phase 2 ran for a total of 21 iterations – 14 balancing and 7 nonbalancing, running for 36 hours on 120 processors on the Cray T3E at NERSC.

5.4. Future Work

The performance of SPGOPS in the CASP competition is very encouraging, but there are several ways in which we believe the algorithm can be improved.

5.4.1. Smoothing the energy function In earlier work the performance of our algorithm has been enhanced by use of smoothing, that is by replacing the objective function by a function that mirrors the coarse grain structure of the original function but that has a smoother fine grain structure. Thus the smoothed function has fewer minimizers and should be easier to minimize globally. In the context of molecular conformation problems, smoothing has been considered by a number of researchers including [6–8,24,25,29,30,34,38,41–43]. We have

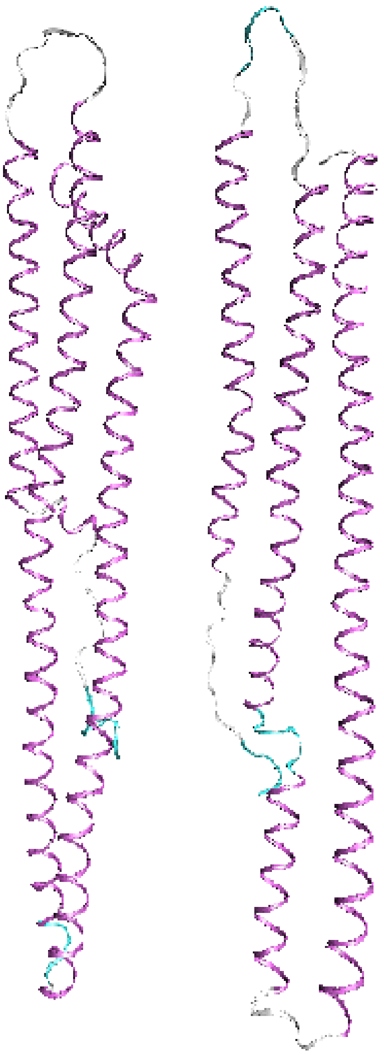

Fig. 7. Results for SPGOPS on Target 124 (containing 242 amino acids) with RMSD of 8.46 Å. The structure to the right is our lowest energy submission, and the structure to the left is the crystal structure.

developed a simple, analytic smoothing function that replaces the terms with the form $1/r$ in the non-bonded portion of equation (1) with $\left(\frac{1+\gamma}{r^2+\gamma}\right)^{1/2}$ where γ is a non-negative constant that is reduced to 0 as the algorithm proceeds. In addition, in the Lennard-Jones term where $1/r$ is raised to the sixth and twelfth power, we instead raise the new term to the P^{th} and $2P^{th}$ power where $P \leq 6$. Using $\gamma > 0$ removes the poles, and using both $P < 6$ and $\gamma > 0$ widen the basins of attraction of local minimizers. In [1,2] we show that the use of this smoothing approach within our stochastic perturbation framework on a simple helical protein results in a lower energy configuration than without, and that the algorithm with smoothing finds low configurations far more efficiently and often than without. We have not yet, however, used smoothing in conjunction with harder proteins such as those discussed in this paper. In the future, we will apply the smoothing approach to some of the more difficult protein targets we have recently been predicting, in the hopes of improving accuracy and/or efficiency.

5.4.2. Improvements to the Global optimization The β biasing function described in Section 4.1 has been recently developed, and its parameters may require some modification to cause hydrogen bonds to form more effectively between β-strands. We will also improve the heuristics for picking dihedral angles to enable the global optimization to focus in areas that will lead to the most diverse new structures and/or greatest overall improvements in energy, at different stages within Phase 2 (sect. 4.2). Finally, our studies of efficiency issues in both the local and global minimizations, indicate that it may be feasible to limit the computational costs of the local minimizations without penalizing the overall accuracy of the method.

6. Summary and Conclusions

We have presented a global optimization approach to the problem of protein structure prediction and discussed some promising results from this approach. We have shown that the SPGOPS approach may be useful for predicting the structures of proteins with new folds that are not found in the sequences or folds of known proteins. Optimization techniques are fundamental both in the large-scale global optimization approach and in the use of biasing or penalty function approaches to encourage the formation of predicted secondary structure. Using an optimization approach to protein structure prediction may also be helpful in terms of developing good energy models for proteins, and understanding the underlying protein folding phenomena.

References

1. A. Azmi, Use of smoothing methods with stochastic perturbation for global optimization (a study in the context of molecular chemistry), Phd thesis, University of Colorado, (1998).
2. A. Azmi, R. Byrd, E. Eskow, R. Schnabel, S. Crivelli, T. Philip and T. Head-Gordon, Predicting Protein Tertiary Structure Using a Global Optimization Algorithm with Smoothing, *Optimization in Computational Chemistry and Molecular Biology: Local and Global Approaches*, C.A. Floudas and P.M. Pardalos, eds., (Kluwer Academic Publishers, 2000) 1-18.
3. P. A. Bates and M. J. E. Sternberg, Model building by comparison at CASP3: Using expert knowledge and computer automation, Proteins: Structure, Function and Genetics **37**, (1999) 47-54.
4. D. F. Burke, C. M. Deane, H. A. Nagarajaram, N. Campillo, M. Martin-Martinez, J. Mendes, F. Molina, J. Perry, B. V. B. Reddy, C. M. Soares, R. E. Steward, M. Williams, M. A. Carrondo, T. L. Blundell, and K. Mizuguchi, An iterative structure-assisted approach to sequence alignment and comparative modeling, Proteins: Structure, Function and Genetics **37**, (1999) 55-60.
5. B. Brooks, R. Bruccoleri, B. Olafson, D. States, S. Swaminathan, and M. Karplus, CHARMM: A program for macromolecular energy, minimization and dynamics calculations, J. Comp. Chem **4**, (1983) 187-217.
6. T. Coleman, D. Shalloway, and Z. Wu, Isotripic effective energy simulated annealing searches for low energy molecular cluster states, Comp. Optim. Appl. **2**, (1993) 145-170.
7. T. Coleman, D. Shalloway, and Z. Wu, A parallel build-up algorithm for global energy minimizations of molecular clusters using effective energy simulated annealing, Technical Report CTC93TR130, Advanced Computing Research Institute, Cornell University, Ithaca N.Y., (1994).
8. T. Coleman and Z. Wu, Parallel continuation-based global optimization for molecular conformation and protein folding, Technical Report CTC-94-TR175, Center for Theory and Simulation in Science and Engineering, Cornell, (1994).
9. W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell and P.A. Kollman, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, J. Am. Chem. Soc. **117**, (1995) 5179-5197.
10. S. Crivelli, B. Bader, R. Byrd, E. Eskow, V. Lamberti, R. Schnabel, and T. Head-Gordon, A physics-based approach to protein structure prediction: CASP4 results, In preparation.
11. S. Crivelli, T. Philip, R. Byrd, E. Eskow, R. Schnabel, R. Yu and T. Head-Gordon, A Global Optimization Strategy for Predicting Protein Tertiary Structure: α-helical Proteins, Computers and Chemistry, conference proceedings for New Trends in Computational Methods for Large Molecular Systems, in press.
12. J. A. Cuff, M. E. Clamp, A. S. Siddiqui, M. Finlay, and G. J. Barton, Jpred: A Consensus Secondary Structure Prediction Server, Bioinformatics **14**, (1998) 892-893.
13. F. S. Domingues, W. A. Koppensteiner, M. Jaritz, A. Prlic, C. Weichberger, M. Wiederstin, H. Floeckner, P. Lacknet, and M. Sippl, Sustained performance of knowledge-based potentials in fold recognition, Proteins: Structure, Function and Genetics **37**, (1999) 112-120.
14. R. L. Dunbrack, Jr., Comparative Modeling of CASP3 targets using PSI-BLAST and SCWRL, Proteins: Structure, Function and Genetics **37**, (1999) 81-87.
15. D. Fischer, Modeling three-dimensional protein structures for amino acid sequences of the CASP3 experiment using sequence-derived predictions, Proteins: Structure, Function and Genetics **37**, (1999) 61-65.
16. T. Head-Gordon and F. H. Stillinger, Predicting polypeptide and protein structures from amino acid sequence: antlion method applied to melittin, Biopolymers **33**, (1993) 293-303.
17. T. Head-Gordon, J. Arrecis and F.H. Stillinger, A strategy for finding classes of minima on a hypersurface: implications for approaches to the protein folding problem, Proc. Natl. Acad. Sci. USA **88**, (1991) 11076-11080.
18. T. Head-Gordon, J. M. Sorenson, A. Pertsemlidis and R. M. Glaeser, Differences in hydration structure near hydrophobic and hydrophilic amino acid side chains, Biophys. J. **73**, (1997) 2106-2115.
19. G. Hura, J. M. Sorenson, R. M. Glaeser and T. Head-Gordon, Solution x-ray scattering as a probe of hydration-dependent structuring of aqueous solutions, Perspectives in Drug Discovery and Design **17**, (1999) 97-118.

20. D. T. Jones, Protein secondary structure prediction based on position-specific scoring matrices, J. Mol. Biol. **292**, (1999) 195-202.
21. D. T. Jones, M. Tress, K. Bryson, and C. Hadley, Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure, Proteins: Structure, Function and Genetics **37**, (1999) 104-111.
22. K. Karplus, C. Barrett, M. Cline, M. Diekhans, L. Grate, and R. Hughey, Predicting protein structure using only sequence information, Proteins: Structure, Function and Genetics **37**, (1999) 121-125.
23. K. K. Koretke, R. B. Russell, R. R. Copley, and A. N. Lupas, Fold recognition using sequence and secondary structure information, Proteins: Structure, Function and Genetics **37**, (1999) 141-148.
24. J. Kostrowicki, L. Piela, B.J. Cherayil and H.A. Scheraga, Performance of the diffusion equation method in searches for optimum structure of clusters of Lennard-Jones atoms, J. Phys. Chem. **95**, (1991) 4113-4119.
25. J. Kostrowicki and H.A. Scheraga, Application of the diffusion equation method for global optimization to oligopeptides, J. Phys. Chem. **96**, (1992) 7442-7449.
26. S. M. Le Grand and K. M. Merz Jr, The application of the genetic algorithm to minimization of potential energy functions, J. Global Opt. **3**, (1993) 49-66.
27. J. Lee, A. Lino, D. R. Ripoll, J. Pillardy, H. A. Scheraga, Calculation of protein conformations by global optimization of a potential energy function, Proteins: Structure, Function and Genetics **37**, (1999) 204-208.
28. A. Liwo, J. Lee, D. R. Ripoll, J. Pillardy, and H. Scheraga, Protein structure prediction by global optimization of a potential energy function, Proc. Natl. Acad. Sci. USA **96**, (1999) 5482-5485.
29. J. J. Moré and Z. Wu Issues in large-scale global molecular optimization, Technical Report MCS-P539-1095, Argonne National Laboratory, Argonne, Illinois, 1996a.
30. J. J. Moré and Z. Wu, Distance geometry optimization for protein structures, Technical Report MCS-P628-1296, Argonne National Laboratory, Argonne, Illinois, 1996b.
31. G. Ne'methy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, and H. A. Scheraga, Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides, J. Phys. Chem. **96**, (1992) 6472-6484.
32. J. Novotny, R.E. Bruccoleri and M. Karplus, An analysis of incorrectly folded protein models. Implications for structure prediction, J. Mol. Biol. **177**, (1984) 787-818.
33. A. Pertsemlidis, A. M. Saxena, A. K. Soper, T. Head-Gordon and R. M. Glaeser, Direct, structural evidence for modified solvent structure within the hydration shell of a hydrophobic amino acid, Proc. Natl. Acad. Sci. **93**, (1996) 10769-10774.
34. J. Pillardy and L. Piela, Molecular dynamics on deformed potential energy hypersurfaces, J. Phys. Chem. **99**, (1995) 11805-11812.
35. B. A. Reva, A. V. Finkelstein and J.Skolnick, What is the probability of a chance prediction of a protein structure with an rmsd of 6 Å ?, Folding and Design **3**, (1998) 141-147.
36. A.H.G. Rinnooy Kan and G.T. Timmer, Stochastic methods for global optimization, American Journal of Mathematical and Management Sciences **4**, (1984) 7-40.
37. A.H.G. Rinnooy Kan. and G.T. Timmer, Global optimization, In ***Handbooks in operations research and management science, volume I : optimization***, G.L. Nemhauser, A.H.J. Rinnooy Kan, and M.J. Todd, eds., (North-Holland, 1989) 631-662.
38. D. Shalloway, Application of the renormalization group to deterministic global minimization of molecular conformation energy functions, Global Optimization **2**, (1992) 281-311.
39. J. M. Sorenson and T. Head-Gordon, The importance of hydration for the kinetics and thermodynamics of protein folding: simplified lattice models, Fold and Design **3**, (1998) 523-534.
40. J. M. Sorenson, G. Hura, A, K, Soper, A. Pertsemlidis and T. Head-Gordon, Determining the role of hydration forces in protein folding, Invited Feature Article for J. Phys. Chem. B **103**, (1999) 5413-5426.
41. F. H. Stillinger, Diffusion smoothing, Phys. Rev. B **32**, (1985) 3134-3141.
42. F. H. Stillinger and T. A. Weber, Nonlinear Optimization Simplified by Hypersurface Deformation, J. Statist. Phys. **52**, (1988) 1429-1445.
43. Z. Wu, The effective energy transformation scheme as a special continuation approach to global optimziation with application to molecular conformation, Technical Report CTC-93-TR143, Center for Theory and Simulation in Science and Engineering, Cornell University, (1993).

44. A.-S. Yang and B. Honig, Sequence to structure alignment in comparative modeling using PrISM, Proteins: Structure, Function and Genetics **37**, (1999) 66-72.
45. R. C. Yu and T. Head-Gordon, Neural network design applied to protein secondary structure prediction, Phys Rev. E. **51**, (1995) 3619-3627.
46. A. Zemla, C. Venclovas, K. Fidelis, and B. Rost, A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment, Proteins: Structure, Function and Genetics **34**, (1999) 220-223.
47. H. Zhu and W. Braun, Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting distance geometry calculations of beta-sheet formation in proteins. Protein Science **8**, (1999) 326-342.